#6

KONINKRIJK DER NEDERLANDEN

Bureau voor de Industriële Eigendom

Hierbij wordt verklaard, dat in Nederland op 25 mei 1999 onder nummer 1012148,

ten name van:

KONINKLIJKE KPN N.V.

te Groningen

een aanvrage om octrooi werd ingediend voor:

"Spraakverwerkend systeem",

en dat de hieraan gehechte stukken overeenstemmen met de oorspronkelijk ingediende stukken.

Rijswijk, 25 januari 2000.

De Directeur van het Bureau voor de Industriële Eigendom,
voor deze,

P.J.C. van den Nieuwenhuijsen.

KINGDOM OF THE (crest) NETHERLANDS

PATENT OFFICE

This certifies that in the Netherlands, on 25 May 1999, a patent application was filed under number 1012148, in the name of:

Koninklijke KPN N.V.

of Groningen

for: "Speech-processing system".
And that the documents attached hereto are in accordance with the documents originally filed.

Rijswijk, 25 January 2000.
On behalf of the Chairman of the Patent Office,

(signature)

(P.J.C. van den Niewenhuijsen)

ABSTRACT

To improve the performance of speech recognition under mobile circumstances, it is customary for speech material to be collected in order to be capable of making more accurate models of the speech. However, with some regularity the error correction is changed by the manufacturer, as a result of which the mismatch between training and reality increases. In addition, transmission errors are currently "taken care of" by including them in the training process, which increases the chance of "garbage in, garbage out". In order to overcome said drawbacks, the information available downstream (1, 2) in the frames on the frame quality (BFI) and the presence of speech (SP), is used to dynamically control the upstream speech recogniser (20). The result is that, of frames presumed incorrect, only the correct part is used, and frames in which no speech was transmitted, but in which there is silence, are ignored by the speech recogniser.

(FIG. 1)

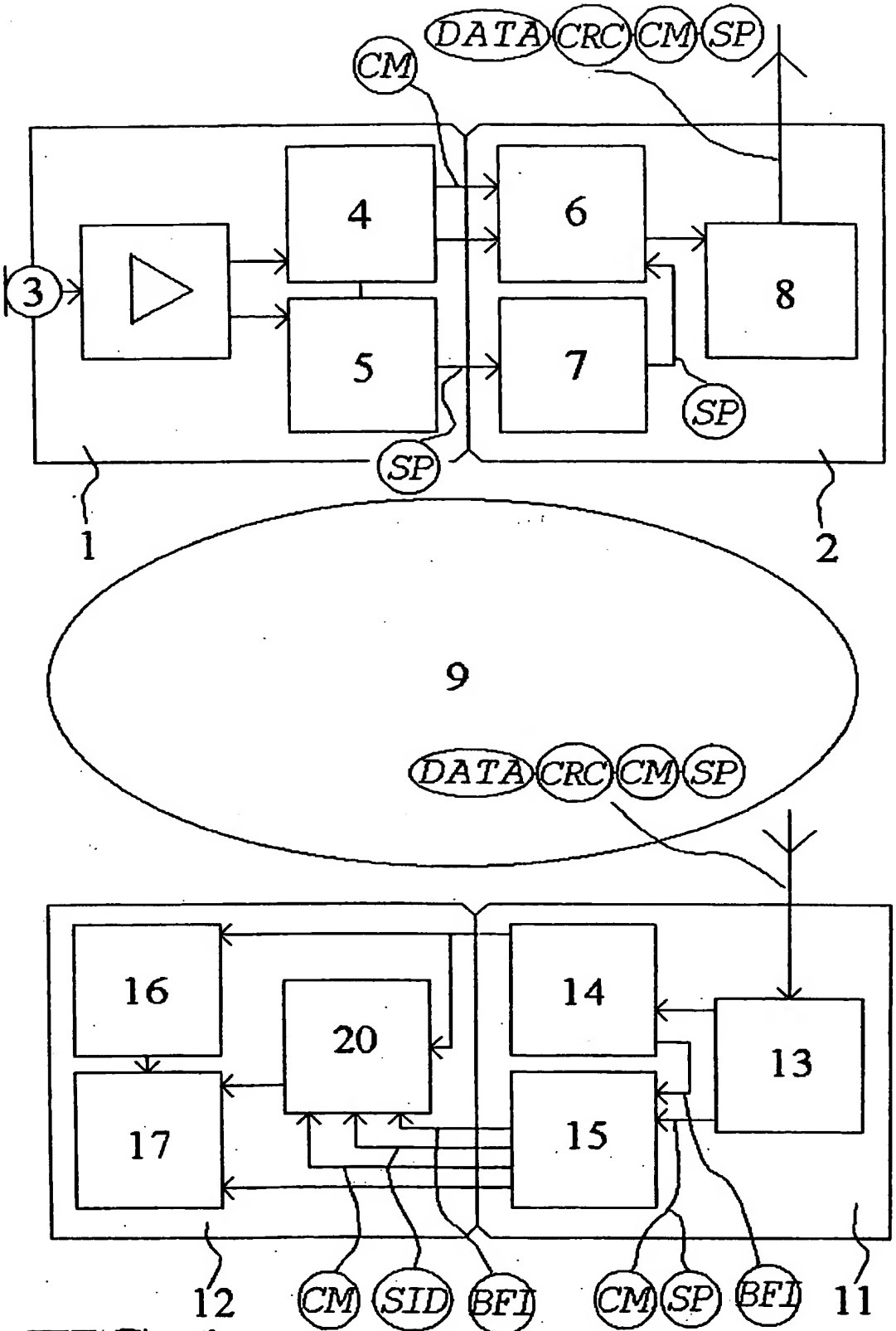
DATA
CRC
CM
SP
CM
4
6
8
3
5
7
SP
SP
1
2
9
DATA
CRC
CM
SP
16
20
14
13
17
15
12
CM
SID
BFI
CM
SP
BFI
11

FIG. 1

Speech-processing system.

BACKGROUND OF THE INVENTION

The invention relates to a speech-processing system, comprising speech-recognition means for processing a signal (DATA) fed from a source to a speech input of said speech-processing system.

It is known that the quality of speech recognition at the receiving side of, e.g., a GSM link [GSM = Global System for Mobile communications] is currently insufficient. If the recogniser is located within the network, the recognition result on the GSM-speech signal received and decoded is partly affected by the amount of artificially generated noise which is added, on the basis of the silence detected at the transmission side, and the noise and disturbances received resulting from the decoded transmission errors on the radio path. To improve recognition, it is customary to collect speech material that had been transmitted, by way of GSM, and to use said material to develop new speech models, which are trained to speech signals containing (artificially generated) noise and distortions due to transmission errors, as a result of which the mismatch between the training situation and the recognition reality may be reduced.

The known matter has the following drawbacks: due to the training on the received and decoded speech signals, the performance of the speech recogniser may only be improved marginally, since:

1) decoding, e.g., encoded GSM signals is not standardised (only encoding is standardised), which signifies that in practice there arise situations in which the speech recogniser is trained on a GSM speech decoder other than the one applied at the input of the recogniser. The error correction applied in the decoder, e.g., is regularly changed since the manufacturer has found an improved way of processing transmission errors (which give rise to damaged speech) in such a manner that a large part of said errors is hidden (and therefore not or hardly noticeable to the human ear). This results in a mismatch arising between the training set on which the speech models are based and the actual speech.

2) by training on speech having transmission errors, one admittedly already models the errors in the speech models (which thereby become more complex), but there is no guarantee that the overall quality of the recognition increases, since there often applies: garbage in, garbage out.

3) it is not known in advance whether a signal contains speech or silence (from the transmitting side). Since artificially generated noise is added at the receiving side (comfort noise) when silences have been observed, the performance of the speech recognition declines, since the recogniser will attempt to "recognise" the noise.

SUMMARY OF THE INVENTION

The object of the invention is to overcome said drawbacks and to improve the performance of automatic speech-recognition systems operating at the receiving side of a speech-frame-oriented telephone-speech link. This may be, e.g., GSM, UMTS [= Universal Mobile Telecommunications System] or Voice-over IP [= Intelligent Peripheral]. The core of the invention is that, at the receiving side, not only a speech signal is offered to the speech-recognition system, but also signal parameters, which give information on characteristics of the signal received.

It concerns, e.g., parameters indicating the presence or absence of speech energy in the signal received, or the realiability of the signal received according to redundancy checks added at the transmitting side (e.g., CRCs [= Cyclic Redundancy Checks]).

In the event of GSM, such parameters are calculated on the basis of frames. Here, the parameters of interest in the framework of the invention are, inter alia, the BFI (= Bad Frame Indicator) calculated from, e.g., the CRC values per frame, and the SID (= SIlence Descriptor) derived from a parameter SP (= Speech Flag). Said parameters are so far only used in GSM for detecting errors in the speech frames received, or for transmitter control (transmit only if speech is present), as the case may be.

Control of a speech recogniser by classifying parameters promotes the accuracy of the recognition, since the artificially generated noise may be ignored and defective frames may either be

ignored or adjusted, e.g., partially processed. Apart from the parameters referred to above – the BFI and the SID – use is also made of an encoding-mode parameter defining the significance of the speech-frame bits (FR [= Frame Relay], EFR [= Enhanced Full Rate], or the various modes under which AMR [= Adaptive Multi Rate] may operate). On the basis hereof, the recognition algorithm operative in the speech recogniser is adjusted to the characteristics with which the speech signal is encoded and decoded.

DESCRIPTION OF THE FIGURES

The operation of the invention is further explained with reference to several figures. As an example, we take the current part of the GSM system which makes use of an Enhanced Full Rate (= EFR) codec[?]. The same does not apply, however, to a Full Rate (= FR) codec, nor to the (future) Adaptive Multi Rate (= AMR) codec. FIG. 1 shows two terminals – a first, mobile terminal, such as a GSM handset, and a second, nonmobile terminal, such as a GSM base station – which are capable of communicating with one another by way of a wireless medium 9. In the figure, there is presented only upstream communication – from handset to base station.

The handset shown in the top part of FIG. 1 comprises two modules or subsystems, namely a TX/DTX Handler 1 (DTX stands for Discontinuous Transmission) and a TX Radio Subsystem 2. Module 1 comprises a microphone 3, a speech encoder 4 and a Voice Activity Detector (= VAD) 5. Module 2 comprises a channel encoder 6, a SPeech-flag monitor 7 and a transmitter 8. Signals received by the microphone 3 are fed to both the speech encoder 4 and to the VAD 5.

In the VAD 5, it is detected whether the microphone 3 is receiving speech or silence. This is encoded with a "SPeech flag" (=SP), which is sent along with each speech frame. In the channel encoder 6, the microphone signal encoded in encoder 4 is encoded into frames capable of being transmitted by way of transmitter 8. To the frames, there is added redundant information, such as a check-sum code (CRC) on the basis whereof it may be calculated, at the receiving side, whether the frame has been transmitted correctly. In specific cases, an

incorrectly transmitted frame may be corrected using said redundant information.

During the setup of the link, it is determined which encoding algorithm is used, which may be represented by the parameter CM (= "coding mode"). In the event of specific speech codecs (e.g., AMR), the "coding-mode" parameter for each frame is sent along, and the recogniser is dynamically driven thereby. In the event of other speech codecs, the parameter is transmitted to the receiving side only once, at the start of a session.

Transmitter 8 thus transmits a frame-encoded signal containing data (the signal proper), the parameter SP, the parameter CM (for specific speech codecs) and redundant information, as contained by the check sum CRC.

The receiving terminal at the bottom of FIG. 1 comprises two modules or subsystems in a GSM base station, namely, an RX[?] Radio System 11 – the counterpart of module 2 of the handset, and an RX DTX Handler 12 – the counterpart of module 1. Module 11 comprises a receiver 13, a channel-decoding and error-correcting module 14 and a parameter detector 15; the latter detects the presence and the value of the parameter SP sent along with the data signal and, if present, the parameter CM. Module 12 comprises a speech decoder 16 and a further processing module 17.

The input of a speech-recognition module 20 is - incidentally, per se in conformity with the prior art - connected to the ouput of the channel decoder 14. The speech recogniser 20 therefore processes the data signal not yet speech-decoded (speech). In conformity with the present invention, the speech recogniser 20 is driven by one or more signal parameters, which are received by way of detector 15. The basis of the parameter SP is formed at the transmitting side in the GSM handset, independently from the signal contents of the data signal received. In the error-correcting module 14, the frames received are checked for correctness, prior to decoding, against the redundant information sent along. Incorrect frames are earmarked as such or, if possible, repaired (in simple cases). Correct frames are passed on to the speech decoder 15. When it is not possible to correct a frame, module 14 gives off a BFI (= Bad Frame Indicator) parameter to detector module 15. According to the invention, said BFI is passed on, apart from to the speech decoder 16, to the speech recogniser 20 as well. Upon receipt of

said BFI, the speech recogniser 20 ignores the input offered, or attempts as yet to recognise that part of the frame which indeed may be earmarked as being correct (although the BFI has been set). In other words, the value of the BFI parameter operates as a control parameter for the speech recogniser, as a result of which it processes only correct frames in one go.

Of frames earmarked as being broken, it is attempted to use only that part which still continues to be correct, and frames earmarked as being wholly incorrect are ignored. That, in the event of a set BFI flag, part of the frame may still be correct, is caused by the bits in the speech frames being broken down into several classes (in GSM: 1A, 1B and 2).

Not every class is "protected" in the same manner by adding redundant information. For, e.g., GSM, if bits of class 1A are characterised as being "damaged" (on the basis of the CRC), the BFI flag is set (some manufacturers also set said flag in the event of damaged 1B bits).

This need not signify, however, that all remaining bits are damaged as well. The recogniser takes, as its input, feature vectors (Rabiner & Juang, 1993). Each speech frame is converted into a feature vector. The values of the undamaged part of the speech frame may still be offered to the recogniser. This may be realised, e.g., by giving the corrupted features in the feature vectors one specific value which results in a nil effect on the score of the signal received (De Veth, Cranen & Boves, 1998), or by ignoring the entire frame (Lippman & Carlson, 1997). In approximately the same way, the SID parameter affects the speech recogniser 20. The SID parameter is derived from the value of the SPeech flag as given off by the Voice Activity Detector 5 and transmitted by transmitter 8. In the event of speech, the SP receives a specific value, as well as the SID; should speech be lacking (silence), the SP and thereby the SID parameter will receive another value. The result is that the speech recogniser is enabled in the event of the transfer of a real speech signal and disabled in the event of the absence of speech. Finally, as indicated above, it is possible to set the operation of the speech recogniser 20 as a function of the encoding algorithm of the speech encoder 4 (e.g., FR, EFR, AMR etc.). In the figure, such is done by the parameter CM determined by way of the

handshake (and therefore during the setup of the link), or sent along with each speech frame.

REFERENCES

Lippmann, R.P., Carlson, B.A., "Missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise", Proc. Of Eurospeech97, Rhodos, Greece, 1997.

Rabiner, L., Juang, B.H., "Fundamentals of Speech Recognition", Prentice-Hall, Inc. New Jersey, 1993.

Veth, J. de, Cranen, B., Boves, L. (1998), "Acoustic backing-off in the local distance computation for robust automatic speech recognition", Proc. Of ICSLP 1998, Sydney, Australia.

CLAIMS

1. Speech-processing system, comprising speech-recognition means (20) for processing a signal entered from a source (1, 2) to a speech input (DATA), characterised by means for affecting the operation of the speech-recognition means by one or more control parameters (CM, SID, BFI) entered by way of a control input, each control parameter relating to a specific characteristic of the signal entered from the source to the speech-recognition means (DATA).

2. Speech-processing system according to claim 1, characterised in that a first control parameter (BFI) relates to the reliability or correctness of the signal entered and that the operation of the speech-recognition means (20) is adjusted to the reliability or correctness, as the case may be, indicated by said first control parameter, of the signal entered.

3. Speech-processing system according to claim 1, characterised in that a second control parameter (SID) relates to the speech/noise ratio and that the operation of the speech-recognition means (20) is adjusted to the speech/noise ratio of the signal entered indicated by said second control parameter.

4. Speech-processing system according to claim 1, the signal entered to the speech-recognition means (20) being encoded in speech-encoding means (4) at the source, characterised in that a third control parameter (CM) relates to the speech-encoding mode in the speech-encoding means, the operation of the speech-recognition means (20) being adjusted to the speech-encoding mode indicated by said third control parameter.

5. Telecommunications system, comprising a first terminal (1, 2) having speech- and channel-encoding means (4, 6), a transmission medium (9) and a second terminal (11, 12) having channel- and speech-decoding means (13, 16) and a speech-processing system according to claim 1, said signal (DATA) being offered from the first terminal, by way of the transmission medium, to the speech input of the speech recogniser of the second terminal, and each control parameter (CM, SID, BFI) being

offered by the first terminal, by way of the transmission medium, to the control input intended for that purpose of the speech-processing system of the second terminal.

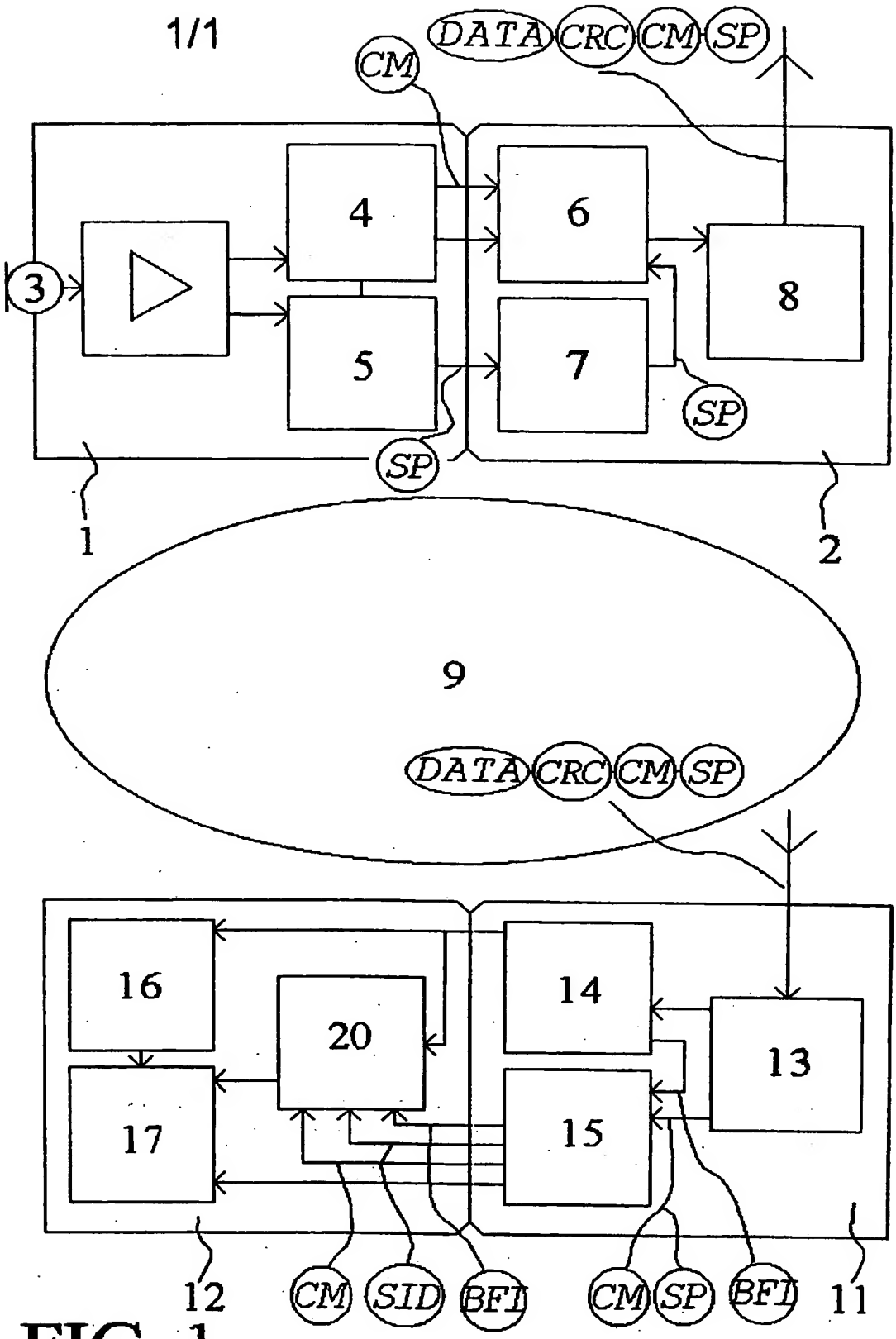
DATA
CRC
CM
SP
CM
4
6
3
8
5
7
SP
SP
1
2
9
DATA
CRC
CM
SP
16
20
14
13
17
15
12
CM
SID
BFI
CM
SP
BFI
11

FIG. 1